Estimation of lineup efficiency effects in Basketball using play-by-play data

L'uso dei dati del play-by-play per la stima degli effetti di quintetto nella pallacanestro

Luca Grassetti, Ruggero Bellio, Giovanni Fonseca and Paolo Vidoni

Abstract The paper aims at defining a data-driven approach to team management in basketball. A model-based strategy, based on a modification of the adjusted plus-minus approach, is proposed for the analyses of the match progress. The main idea is to define a model based on the 5-man lineups instead of the single players. In this framework, given the large number of possible lineups, the regularization issue is quite relevant. The empirical application is based on the data of the current Italian championship (Serie A1). The play-by-play data are considered along with some information resulting from the game box scores.

Abstract *Questo contributo propone un approccio "data driven" al processo di gestione delle squadre nella pallacanestro. In particolare, si sviluppa una strategia "model-based" per l'analisi della progressione delle partite, basata su una versione modificata del modello per l'*adjusted plus-minus*. L'obiettivo è definire il modello sulla base dei quintetti anzichè dei singoli giocatori. In questo contesto, data l'elevata numerosità dei possibili quintetti, il ricorso ai metodi di regolarizzazione risulta essere molto rilevante. L'analisi empirica è basata sui dati dell'attuale campionato della Lega Basket Italiana (Serie A1). Il play-by-play delle partite è considerato congiuntamente ad alcuni dati raccolti dai box score delle partite.*

Key words: Basketball Analytics, Data-driven decision process, Play-by-Play data, Statistical Model, Web-crawling.

Luca Grassetti
Department of Economics and Statistics, University of Udine, Italy
Via Tomadini, 30/A - 33100 Udine (UD) e-mail: luca.grassetti@uniud.it

Ruggero Bellio
e-mail: ruggero.bellio@uniud.it

Giovanni Fonseca
e-mail: giovanni.fonseca@uniud.it

Paolo Vidoni
e-mail: paolo.vidoni@uniud.it

1 Introduction

Sports analytics for team games has principally two main purposes: match outcome prediction and analysis of performance. Even if the two issues are related, methods adopted may be quite different. Basketball makes no exception and outcome prediction literature is based on some model assumption on the final outcome. For example, [16] and [17] assume Poisson distribution for each team final match score in order to compute match-winning probabilities, as usually done in soccer literature. Further extensions are provided by [8] that uses logistic regression in a Markov Chain setting and by [11] that introduces a time variation for the strength of teams in the model. On the other side, there are results based on classification and automatic algorithms, as in [20] and [21] where the authors adopt a machine learning approach in order to predict match results.

Players performance analyses are usually based on box score statistics that are widely available for basketball matches. Moreover, nowadays, data are collected on a real-time basis during the game and possession outcomes are used for describing a single player's performance. For example, [2] relates the individual player contribution to the match-winning probability of the team at different game moments.

A remarkable exception is the estimation of player performances based on the so-called adjusted plus-minus (APM) method. First introduced in an influential contribution by [15], the computation of the APM is based on play-by-play data aggregated in *shifts*. A shift is defined as a period of playing time without any substitution for either team. The APM is computed by considering a linear regression model for the point differential during each shift as the response variable, with regressors given by signed dummy variables for every single player involved in the shift. The estimation is carried out from the perspective of the home team, with the consequential definition of the point difference and the sign of player dummy variables.

The APM method has a strong appeal since the estimated coefficients can be interpreted as (net) player efficiency measures, i.e. they are adjusted for the other players on the field. Not surprisingly, the measure was readily adopted by NBA data analysts. At the same time, the original method is prone to difficulties, since it entails sparse design matrices and multicollinearity. In other words, it is a typical setting where some form of regularization is called for, and, indeed, the Regularized APM (RAPM) was proposed in [18]. The RAPM method employs ridge regression for the estimation of player efficiency and, as summarised in [4], it is more accurate, robust and stable than the original APM. The method was adopted also for the analysis of the players of the Major League Soccer [7] and the National Hockey League (see [9] and [10]).

The proposal of this paper builds upon the RAPM setting, with three important variations. The main innovation is that instead of focusing on the performances of individual players, we focus on the performances of the entire lineups (5-man units) on the field, with the idea that player performances may depend on the interaction with teammates and on the contrastive action of the lineup of the opposite team on the field. From a statistical perspective, regularization is even more essential for the estimation of lineup efficiency than for that of individual players, so that

we may consider also other approaches next to ridge regression, such as empirical Bayes and boosting (see, for instance, [3]). A final distinctive feature of the current proposal is the adoption of a performance index rating as the response variable, thus considering a more comprehensive measure than point differential. The overall aim is to provide a useful tool for coaches that highlights the strength and the weakness of the different available lineups, offering some added value with respect to the evaluation of the performance of individual players.

The paper is organized as follows. Section 2 illustrates the data used for the analysis, obtained from the Italian Basketball League. Since the data wrangling process presents some features that may be of interest for statisticians, it is illustrated with some details. Section 3 introduces the model adopted for the estimation of lineup efficiency, and it reports some results for the case study of interest. Finally, Section 4 contains a brief discussion and some concluding remarks.

2 Data wrangling and data exploration

The empirical analysis described in this paper is based on a dataset regarding the Italian Basketball League (Serie A1). In particular, the matches of the first round of the current championship 2018/2019 are considered. The dataset collects the play-by-play information along with the matches box scores, which are made available by the league website (www.legabasket.it). The plays are then aggregated in shifts, which is the aggregation level considered in the statistical analysis. In this section, the data wrangling process is described, emphasizing some key aspects which could be interesting from the statistical perspective, and the results of a preliminary data exploration are briefly presented.

2.1 Data collection

In order to collect the data from the Italian Basketball League website, we use the R statistical software [13]. In particular, for performing the data scraping process we consider the `rvest` package [19]. The `scrapeR` [1] and `Rcrawler` [6] packages may also be used to this end. For every single match, we collect both the box scores data and the play-by-play information. A play is defined as an event during the possession involving a positive or negative evaluation (see Table 1).

For the collection of the box scores data of a single match, the associated specific web page is parsed, the tables are then identified in the text using the function `html_nodes` and, finally, the box scores information are organized in a data frame with the function `html_table`.

The play-by-play data can be obtained using a similar procedure, but some useful tricks have to be considered in order to produce a ready-to-use dataset. To this end, the information available in the play-by-play table is collected as raw text extracted

using the command `html_text`. Furthermore, the textual output is pre-processed using the `stringi` [5] package, with the aim to code the information in a usable way. Each play is finally recorded by considering the following features:

- player finalizing the play,
- possible intermediate events in a play (substitutions, time-outs and so on),
- outcome of the play (classified according to different potential categories),
- quarter,
- minute in the quarter,
- home and the away teams,
- team of the finalizing player.

These data are then completed using the box scores tables. In particular, by considering the information on the starting five, we are able to reconstruct the 5-man unit involved in each play. In fact, whenever the event in the play-by-play dataset is a substitution, the change in the 5-man unit is recorded. This piece of information is crucial in order to aggregate the plays in shifts, as required for subsequent analyses.

2.2 Data cleansing

The data cleansing process is a further important step in order to define the dataset in the required form. At first, a check on the names of the involved players is required, since they are sometimes reported with errors. Furthermore, the players having an average play time shorter than five minutes are removed from the analysis, so that the 5-man units involving these players will present one or more anonymized individuals, as proposed in the APM literature (see for example [4]). We call these individuals *dummy players*. The original lineups are also recorded.

A numerical variable is then defined by some specific outcomes produced in the plays. In particular, the scores reported in Table 1 are assigned to the offensive team, and opposite scores are assigned to the defensive team. The scores are defined by considering only those events deemed as the most relevant for the outcome of the play.

Table 1 Scores of the events used in the computation of the outcome measure for each play.

Value	Events
-1	missed free-throw, turnover or offensive foul
-0.5	missed shot (2 points or three points shots)
0.5	assist
1	steal, offensive or defensive rebound, block, scored free-throw or received foul
2	scored shot
3	scored three-pointer

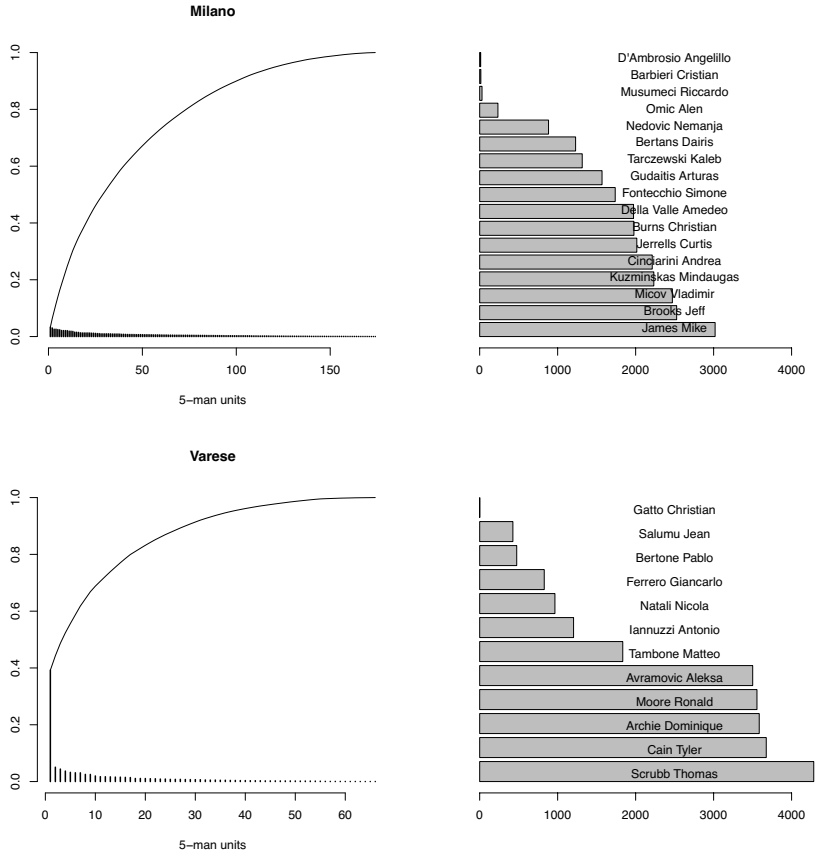
The records regarding turnovers caused by steals and personal defensive foul corresponding to a received foul are removed in order to prevent duplicated infor-

mation. Assists could also be removed from the analysis if they are deemed as not particularly relevant. The sum of the above single scores produces a result which is not far from the final result of the match, and it has the merit of accounting for some non-scoring key events.

2.3 Some summary statistics

The data used for the analysis concern 120 matches, resulting in 4108 shifts with more than 41000 plays, corresponding to around 21000 possessions. The total number of lineups is around 2000. The latter number would be actually slightly larger had we not replaced each player with low playing time per game by a single team-specific dummy player.

Fig. 1 Distribution of the number of plays for Milano and Varese team.



The left panels of Figure 1 reports the distribution of the number of plays for lineups for two selected teams, Milano and Varese, that are the two teams with the largest and smallest number of lineups, respectively. The right panels of the same figure report the total number of plays for the players with the highest totals.

3 Estimation of lineup effects

The starting point for the estimation of lineup effects is a simple model for the score of the t shift, with $t = 1, \ldots, T$,

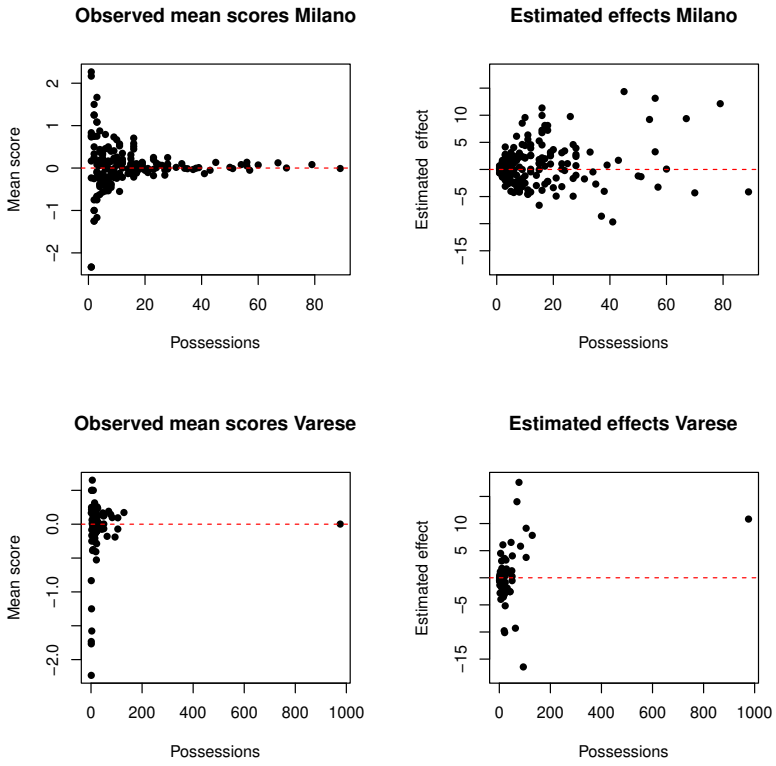
$$y_t = \beta_0 + \mu_{h[t]} - \mu_{a[t]} + \varepsilon_t \,. \tag{1}$$

Here we consider the entire data set encompassing all the available matches so that T=4108. The response variable in (1) is given by the difference between the mean outcome for the home team and the mean outcome of the away time for each shift, where the mean is over the number of possessions for each team. In case only one team is involved in a given shift, the mean outcome of the opposite team is replaced by the grand mean over the entire sample, along the lines of what done in the APM literature. The notation $h[t]$ and $a[t]$ identifies the lineup for the home and away team for shift t, respectively. More precisely, $h[t]$ and $a[t]$ assume a value in the set $1, \ldots, N$, where N is the total number of lineups in the dataset ($N = 1998$ in the application). The model specification could be made more complex by including some covariate effects, but for the sake of illustration, we focus here on a basic form.

The estimation of the lineup effects μ is based on regularized weighted linear regression, with weights corresponding to the total number of possessions of every shift. Here we report some results corresponding to the estimation based on an empirical Bayes approach, assuming normal lineup effects. The estimation has been carried out by means of the `hglm` R package [14], which allows for the inclusion of observation weights and estimates the regularization parameter by REML. Note that the results obtained with ridge regression were essentially equivalent to those obtained by empirical Bayes, but the latter approach seems preferable since it allows for straightforward inclusion of further regression covariates. Instead, the results based on boosting were somewhat unsatisfactory, exhibiting an apparent lack of shrinkage and requiring much longer computations than the two other methods.

The left panels of Figure 2 report the mean response for each lineup of the same two teams of Figure 1 against the number of possessions. The right panels of Figure 2 report instead the estimated effects $\widehat{\mu}$ for the lineups of the two teams against the number of possessions. The effects are on the 100-possession scale, as customary in the APM literature. It is apparent that the estimated lineup effects adjust for the quality of the other lineups on the field. A remarkable instance is the lineup with the highest number of possessions for the Varese team, for which the apparent null observed mean score corresponds to a positive estimated effect. Indeed, this lineup has started most of the team matches, thus playing against higher-quality lineups.

Fig. 2 Observed mean scores against number of possessions (left), and lineup estimated effects against number of possessions (right), for Milano and Varese team.



4 Conclusion and ongoing research

Play-by-play data represent an invaluable source of information for the statistical analysis of basketball results. This paper has illustrated how to obtain the data and perform some pre-processing using R software tools, which may be familiar to many statisticians. The analysis reported is an extension of the RAPM approach to the analysis of lineup effects, with the important difference of considering a more comprehensive outcome measure.

Estimates of lineup effects should be of interest for team managers and data analysts, since they provide some insight about the team lineup strategy adopted. The quantification of lineup performances extends the information provided by player-based RAPM analysis. Ongoing research concerns the study of the connection existing between the two approaches, as well as the possibility of disentangling both sets of measures into more than one dimension corresponding to different aspects of the game, thus exploiting even further the availability of play-by-play data.

Acknowledgments

We are grateful to the Italian Basketball League for the permission of using the play-by-play data. This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003.

References

1. Acton, R.M.: scrapeR: Tools for scraping data from HTML and XML documents. R package version 0.1.6. https://CRAN.R-project.org/package=scrapeR (2010)
2. Deshpande, S.K. and Jensen, S.T.: Estimating an NBA player's impact on his team's chances of winning. J. Quant. Anal. Sports, **12**, 51–72 (2016)
3. Efron, B. and Hastie, T.: Computer Age Statistical Inference. Cambridge University Press, Cambridge (2016)
4. Engelmann, J.: Possession-based player performance analysis in basketball (adjusted +/- and related concepts). In: Handbook of Statistical Methods and Analyses in Sports, pp. 231-244. Chapman and Hall/CRC (2017)
5. Gagolewski M. *et al*: R package stringi: character string processing facilities. http://www.gagolewski.com/software/stringi/ (2019)
6. Khalil, S.: Rcrawler: web crawler and scraper. R package version 0.1.9-1. https://CRAN.R-project.org/package=Rcrawler (2018)
7. Kharrat, T., Pena, J.L. and McHale, I.: Plus-minus player ratings for soccer. arXiv preprint arXiv:1706.04943, (2017)
8. Kvam, P. and Sokol, J.S.: A logistic regression/Markov chain model for NCAA basketball. Naval Research Logistics (2006)
9. Macdonald, B.: A regression-based adjusted plus-minus statistic for NHL players. Journal of Quantitative Analysis in Sports **7.3**, (2011)
10. Macdonald, B.: Adjusted plus-minus for NHL players using ridge regression with goals, shots, fenwick, and corsi. Journal of Quantitative Analysis in Sports **8.3** (2012)
11. Manner, H.: Modeling and forecasting the outcomes of NBA basketball games. Journal of Quantitative Analysis in Sports, **12**, 31–41 (2016)
12. Omidiran, D.: A new look at adjusted plus/minus for basketball analysis. MIT Sloan Sports Analytics Conference [online], (2011)
13. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
14. Ronnegard, L., Shen, X. and Alam, M.:hglm: A package for fitting hierarchical generalized linear models. The R Journal **2**: 20–28 (2010)
15. Rosenbaum, D.: Measuring how NBA players help their teams win. Retrieved from http://www.82games.com/comm30.htm (2004)
16. Ruiz, F.J.R. and Perez-Cruz, F.: A generative model for predicting outcomes in college basketball. J. Quant. Anal. Sports, **11**, 39–52 (2015)
17. Shen, K.: Data analysis of basketball game performance based on bivariate poisson regression model. Computer Modelling & New Technologies, **18**, 474–479 (2014)
18. Sill, J.: Improved NBA adjusted +/- using regularization and out-of-sample testing. In: Proceedings of the 2010 MIT Sloan Sports Analytics Conference (2010)
19. Wickham H.: rvest: easily harvest (scrape) web pages. R package version 0.3.2. https://CRAN.R-project.org/package=rvest (2016)
20. Yang, J.B. and Lu, C.-H.: Predicting NBA Championship by learning from history data. Proceedings of Artificial Intelligence and Machine Learning for Engineering Design (2012)
21. Zimmermann, A., Moorthy, S. and Shi, Z.: Predicting NCAAB match outcomes using ML techniques - Some results and lessons learned. Proceedings ECML 2013, 69–78 (2013)